APPROXIMATE CONVOLUTION OF THE GAMMA AND MIXED GAMMA DISTRIBUTIONS

H. C. S. THOM

Environmental Data Service, ESSA, Silver Spring, Md.

ABSTRACT

Approximate convolution methods are developed for unmixed and mixed precipitation distributions. Results of applying these to monthly distributions to obtain the distributions of sums of various numbers of months are presented. These are compared to distributions fitted to the sums by employing the confidence intervals for the fitted distributions.

1. INTRODUCTION

The unmixed gamma distribution has been used extensively in fitting precipitation data. After the author [1] proposed a mixed distribution for precipitation series where zeros occur this distribution also began to be widely employed. Since it is often necessary to add precipitation, i.e. June plus July, for example, it is a great convenience to be able to accomplish this analytically and so avoid tabulating and fitting the summed series. The approximate convolution of two unmixed precipitation distributions was first given in [2]. Since then the approximation has been tested many times and extended to the sum of any number of variates. More recently, an approximate convolution of the mixed gamma distribution has been developed. Both approximations are discussed here.

2. THE GAMMA DISTRIBUTION

The gamma probability density function is

$$f(x;\beta,\gamma)=\frac{1}{\beta^{\gamma}\Gamma(\gamma)}\,x^{\gamma-1}e^{-x/\beta}. \tag{1}$$

The requirements for the convolution of several gamma distributions are that the variates be independent and have the same scale. Hence, the sum of several independent gamma variates with scale β and shapes γ_1, and $\gamma_2, \ldots \gamma_i, \ldots \gamma_n$ each having the distribution $f(x_i;\ \beta,\ \gamma_i)$ have the distribution of the sum $f\left(\sum_{i=1}^{m} x_i;\beta,\ \sum_{i=1}^{m}\gamma_i\right)$. Thus, the gamma distribution is one of a special family of distributions which repeats itself in convolution. Unfortunately, these stringent conditions are not often met exactly in climatological series. The great advantage of being able to form convolutions, however, makes any approximate method very useful.

The first three moment's of the gamma distribution will be needed. These are given by the following equations for the first three moments [3]:

$$\mu=\beta\,\gamma, \tag{2}$$

$$\sigma^2=\beta^2\gamma, \tag{3}$$

and

$$\mu_3=2\beta^3\gamma. \tag{4}$$

From (3) and (4) it follows that the skewness statistic is

$$\sqrt{b_1}=2/\sqrt{g}. \tag{5}$$

This will be employed in assessing the accuracy of the approximations of the distribution statistics.

Since the shape parameters add in convolution of gamma distributions, and the mean $\bar{x}$ is also a sufficient estimator of μ it seemed natural to sum the shape statistics and determine the scale statistic from the summed mean. This gives the following two formulas for g estimating γ and b estimating β:

$$g_{12\ldots m}=\sum_{i=1}^{m} g_i \tag{6}$$

and

$$b_{12\ldots m}=\sum_{i=1}^{m}\bar{x}_i/g_{12\ldots m}. \tag{7}$$

Birmingham, Ala., and Columbia, Mo., were chosen at random for application of formulas (6) and (7). The results are shown in table 1, the estimated values and the approximated values being obtained by fitting the summed series and the components of the approximated values by maximum likelihood. The approximations for g appear to be somewhat divergent; however, b and g are negatively correlated and, therefore, tend to compensate each other. In nonreal time prediction it is the quantiles which are of main interest. The 0.10 and 0.90 quantiles are often of practical use there and can be readily obtained from the tables of [4]. These are compared in table 2. The approximations here are quite acceptable being well within the sampling errors.

The gamma distribution approaches the normal distribution as γ becomes large. Although this approach is mathematically slow it is practically fast enough so

TABLE 1.—*Approximated and estimated values of g and b*

	Approximated		Estimated	
	g	*b*	*g*	*b*
Birmingham, Ala.				
June+July+Aug.	9.85	1.37	10.70	1.26
Annual	39.73	1.34	30.98	1.72
Columbia, Mo.				
Annual	32.93	1.21	35.20	1.13

TABLE 2.—*Approximated and estimated values of 0.10 and 0.90 quantiles*

	Approximated		Estimated		$\sqrt{b_1}$
	0.10	*0.90*	*0.10*	*0.90*	
Birmingham, Ala.					
June+July+Aug.	8.36	19.21	8.55	18.96	0.61
Annual	42.77	64.34	41.34	65.71	0.36
Columbia, Mo.					
Annual	31.14	48.81	31.41	48.49	0.34

that annual precipitation is almost always close to normally distributed if it is large enough, and in wetter regions the total of several months may also be close to normal. For example, from table 1 it is seen that the estimated g for June+July+August for Birmingham is 10.70. Since the tables of [4] have the argument t which is scaled in β, or in effect β is unity, and by equations (2) and (3) $\mu=\gamma$ and $\sigma=\sqrt{\gamma}$, we may express the normal deviate by

$$z=(t-\gamma)/\sqrt{\gamma}. \tag{8}$$

In the direct gamma tables [4] we find by interpolation $t(0.10,\ 10.70)=6.7802$, and $t(0.90,\ 10.70)=15.0472$. Substituting these and $\sqrt{g}=3.2711$ in (8) gives

$$z(0.10)=(6.7802-10.70)/3.2711=-1.198$$

and

$$z(0.90)=(15.0472-10.70)/3.2711=1.329.$$

Referring to a standard normal distribution integral table we find

$$N(-1.198)=0.115$$

and

$$N(1.329)=0.908.$$

These values compare with 0.10 and 0.90 on the gamma distribution. Thus, the normal approximation is only off 0.015 probability at the 0.10 quantile and 0.008 at the 0.90 quantile for a 3-mo. total precipitation series.

A similar check may be made more quickly using the skewness $\sqrt{b_1}=k$ in Tolley's [5] table of the first two terms of the Edgeworth series which approximates the gamma distribution closely. This is shown by interpolating on the skewness to obtain $\sqrt{b_1}=0.61$ and $z(0.10)=-1.198$ and $z(0.90)=1.329$. This gave $W(-1.198,\ 0.61)=0.106$ and $W(1.329,\ 0.61)=0.895$ where W is the Edgeworth distribution function. These agree closely with 0.115 and 0.908 obtained above using the gamma distribution. To use Tolley's table it is only necessary to enter them with the values obtained from equation (8) and $\sqrt{b_1}$ as was done above as a check. Thus, for a total of June+July+August at Birmingham the Edgeworth approximation is only off by about 0.01 at both $F=0.10$ and $F=0.90$, a very good approximation for practical purposes. Clearly for larger g or smaller $\sqrt{b_1}$ the approximation will be even closer.

3. THE MIXED GAMMA DISTRIBUTION

The mixed gamma distribution as it has been called here is a mixture of zeros and nonzero values, the latter values distributed approximately as a gamma variate. It arises from two different physical populations, one with zero values and the other with nonzero values. The main application has been to precipitation where two general physical systems operate—one producing precipitation the other not producing it. Its distribution function, i.e., the probability of being less than x, may be expressed by the equation

$$G(x)=q+pF(x) \tag{9}$$

where $F(x)$ is the gamma distribution function or integral from zero to x of equation (1), q is the probability of zero precipitation, and $p=1-q$. The formal convolution of this distribution is complicated and not useful in applications. Later work produced an alternative theoretical model but it is not practical because of difficulties in estimating the parameters. However, the distribution (9) has been employed extensively with good results possibly because of the good estimate q usually provides; hence, some approximate convolution should be very useful.

It is clearly seen that the convolution of (9) would entail products in which q would become smaller as it should since the longer the time period of the variate the less is the probability of a zero. It was, therefore, observed that the convolution of mixed distributions could be put in the form

$$G\left(\sum_{i=1}^{m}x_i\right)=q_1q_2\ldots q_m+(1-q_1q_2\ldots q_m)F\left(\sum_{i=1}^{m}x_i\right) \tag{10}$$

where $F\left(\sum_{i=1}^{m}x_t\right)$ is not a gamma distribution but is only closely approximated by one and hence the shape parameters do not add. It was found, however, that convolution could be accomplished by summing the variances of the individual distributions which are known from equation (3) and the zero probability q to obtain the variance of the sum taking advantage of the fact that the correlation between precipitation events is usually very low and, therefore, there is no appreciable covariance.

The development is most easily carried out for two distributions and then is easily extended to many distributions. If x_1 is the continuous variate with k sample values and x_{01} is the mixed variate with n values then the sample variance of a single distribution is given by

$$\begin{aligned} v(x_{01})&=[\sum_{j=1}^{k}(x_{1j}-\bar{x}_1)^2+(n-k)(0-\bar{x}_1)^2]/n \\ &=p_1v(x_1)+q_1\bar{x}_1^2 \end{aligned} \tag{11}$$

where the summation in the first equality is on sample, not on variables. The mean of x_{01} is

$$\bar{x}_{01}=p_1\bar{x}_1. \tag{12}$$

Substituting the available parameter estimates b_1 and g_1 yields

$$v(x_{01})=p_1b_1^2g_1+q_1b_1^2g_1^2 \tag{13}$$

and

$$\bar{x}_{01}=p_1b_1g_1. \tag{14}$$

The general principle to be followed is to add the variances of the mixed variates to obtain the variance of the mixed convolved variate. Then remove the effect of the zeros to obtain the variance of the approximated gamma variate and add the probability of zeros in the convolved variate to obtain the final mixed distribution. It was found to be more simple to first accomplish this for two variates after which it is easily extended to any number of variates. Using equations (13) and (14) for variates x_{01} and x_{02} and adding gives

$$v(x_{01}+x_{02})=p_1b_1^2g_1+q_1b_1^2g_1^2+p_2b_2^2g_2+q_2b_2^2g_2^2 \tag{15}$$

and

$$\overline{(x_{01}+x_{02})}=p_1b_1g_1+p_2b_2g_2, \tag{16}$$

the variance and mean of the mixed convolution. Now the variance of the mixed convolution is also directly from the mixture

$$v(x_{01}+x_{02})=(1-q_1q_2)v(x_1+x_2)+q_1q_2\overline{(x_1+x_2)}^2. \tag{17}$$

Solving for $v(x_1+x_2)$ yields

$$v(x_1+x_2)=[v(x_{01}+x_{02})-q_1q_2\overline{(x_1+x_2)}^2]/(1-q_1q_2). \tag{18}$$

The mean of $(x_{01}+x_{02})$ by (12) is given by

$$\overline{(x_{01}+x_{02})}=(1-q_1q_2)\overline{(x_1+x_2)} \tag{19}$$

which solving for the unmixed mean gives

$$\overline{(x_1+x_2)}=\overline{(x_{01}+x_{02})}/(1-q_1q_2). \tag{20}$$

Substituting the parameter estimates yields

$$\overline{(x_1+x_2)}=(p_1b_1g_1+p_2b_2g_2)/(1-q_1q_2). \tag{21}$$

Squaring (19) and substituting in (17) yields

$$v(x_1+x_2)=\frac{v(x_{01}+x_{02})}{1-q_1q_2}-\frac{q_1q_2\overline{(x_{01}+x_{02})}^2}{(1-q_1q_2)^3}. \tag{22}$$

Substitution of the distribution statistics from equations (15) and (16) gives finally

$$v(x_1+x_2)=\frac{p_1b_1^2g_1+q_1b_1^2g_1^2+p_2b_2^2g_2+q_2b_2^2g_2^2}{1-q_1q_2}-\frac{q_1q_2(p_1b_1g_1+p_2b_2g_2)^2}{(1-q_1q_2)^3}. \tag{23}$$

Since by equations (2) and (3)

$$b_{012}=v(x_1+x_2)/\overline{(x_1+x_2)} \tag{24}$$

and by (2)

$$g_{012}=\overline{(x_1+x_2)}/b_{012}. \tag{25}$$

The mixed convolution of the distributions becomes

$$G(x_{01}+x_{02})=q_1q_2+(1-q_1q_2)[F(x_1+x_2;\ b_{012},\ g_{012})]. \tag{26}$$

The key equations are (21), (23), (24), and (25). These may readily be extended to m variates by simply introducing the summation and product notations as follows:

Equation (21) becomes

$$\overline{\sum_{i=1}^{m}x_i}=\left(\sum_{i=1}^{m}p_ib_ig_i\right)\Big/\left(1-\prod_{i=1}^{m}q_i\right), \tag{27}$$

equation (23) becomes

$$v\left(\sum_{i=1}^{m}x_i\right)=\frac{\sum_{i=1}^{m}(p_ib_i^2g_i+q_ib_i^2g_i^2)}{1-\prod_{i=1}^{m}q_i}-\frac{\left(\sum_{i=1}^{m}p_ib_ig_i\right)^2\prod_{i=1}^{m}q_i}{(1-\prod_{i=1}^{m}q_i)^3}, \tag{28}$$

equation (24) becomes

$$b_{012\ldots m}=v\left(\sum_{i=1}^{m}x_i\right)\Big/\overline{\sum_{i=1}^{m}x_i}, \tag{29}$$

and equation (25) becomes

$$g_{012\ldots m}=\overline{\sum_{i=1}^{m}x_i}/b_{012\ldots m}. \tag{30}$$

These equations were tested on a series of mixtures shown in table 3 where the months between and including

TABLE 3.—*Results of tests on a series of mixtures for three selected stations*

City & State	Months	Estimated		Approximated		$q_1 \ldots q_m$	Estimated		Approximated	
		b	g	b	g		0.10	0.90	0.10	0.90
Phoenix, Ariz.	I-II	0.80	1.88	0.93	1.54	0.003	0.37	2.95	0.29	2.97
Do.	IV-V	0.84	0.69	0.66	0.89	0.092	0.03	1.45	0.05	1.39
Do.	V-VI	0.32	0.80	0.38	0.73	0.228	0.02	0.62	0.01	0.69
Do.	IX-X	0.75	1.75	0.89	1.55	0.017	0.30	2.63	0.28	2.86
Do.	X-XI	0.66	1.43	0.84	1.12	0.039	0.17	1.99	0.12	2.11
Do.	XI-XII	1.01	1.49	0.98	1.57	0.030	0.29	3.15	0.32	3.17
Do.	IX-XI	0.84	2.19	0.96	1.94	0.004	0.53	3.49	0.48	3.65
Do.	X-XII	1.05	1.80	0.92	2.07	0.005	0.45	3.76	0.52	3.67
Do.	I-XII	1.27	5.60	0.88	8.13	0.000	3.63	11.13	4.18	10.50
Lander, Wyo.	X-XI	0.90	2.66	1.09	2.20	0.000	0.81	4.37	0.70	4.56
Do.	XI-XII	0.51	2.67	0.68	2.01	0.002	0.46	2.49	0.37	2.66
Do.	X-XII	0.77	3.76	0.97	2.98	0.000	1.22	4.89	1.06	5.13
Red Bluff, Calif.	V-VI	0.99	1.13	1.03	1.51	0.007	0.15	2.50	0.31	3.24
Do.	VII-VIII	0.20	0.87	0.48	0.37	0.616	0.01	0.41	0.00	0.51
Do.	IX-X	1.55	1.22	1.43	1.29	0.011	0.28	4.14	0.17	3.99
Do.	V-X	1.31	2.26	1.26	2.74	0.000	0.88	5.62	1.19	6.25

the Roman numerals indicate the period added. The stations were chosen for their relatively large q values during certain months in order to make the tests stringent. The estimated statistics were obtained by fitting the actual summed data series—the approximated statistics from the individual monthly distribution statistics including the mixture statistic q. Only the statistics for the nonzero part of the distribution were tested as it did not appear necessary to test the final q statistic as these have generally shown good agreement with those of the summed series. Also its components had already entered into the determination of the continuous distribution.

The parameter estimates again appear to compensate each other due to the negative correlation between b and g. The approximated 0.10 and 0.90 quantiles appear to show remarkable agreement with the estimated values. This agreement may be evaluated using confidence intervals on the estimated 0.10 and 0.90 probabilities of table 3. The 0.90 confidence intervals (binomial) are appropriate for meteorological work, and these may be found in Dixon and Massey [6] for the 0.10 and 0.90 probabilities and converted to t by interpolation in the inverse tables of [4] and multiplying by the scale b. Of the 32 values tested only two fell slightly out of the confidence intervals. With the 0.90 interval three could have been expected to fall out. The departures of the approximation from the parameter estimates may therefore be considered to be random.

ACKNOWLEDGMENTS

The author is grateful to Marcella D. Thom for programming and computing the approximations, parameter fits, and quantiles and to Maurice Kasinoff and Denaire S. Pyle for tabulating the basic data and computing the confidence limits.

REFERENCES

1. H. C. S. Thom, "A Frequency Distribution for Precipitation," Abstract, *Bulletin of the American Meteorological Society*, Vol. 32, No. 10, Dec. 1951, p. 397.
2. H. C. S. Thom, "A Statistical Method of Evaluating Augmentation of Precipitation of Cloud Seeding," *Technical Report* No. 1, U.S. Advisory Committee on Weather Control June 1957, 62 pp. (see p. 11).
3. H. C. S. Thom, "A Note on the Gamma Distribution," *Monthly Weather Review*, Vol. 86, No. 4, Apr. 1958, pp. 117–122.
4. H. C. S. Thom, "Direct and Inverse Tables of the Gamma Distribution," *ESSA Technical Report* Eds 2, Environmental Data Service, Apr. 1968, 30 pp.
5. H. R. Tolley, "Frequency Curves of Climatic Phenomena," *Monthly Weather Review*, Vol. 44, No. 11, Nov. 1916, pp. 634–642.
6. W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis*, 2d Edition, McGraw-Hill Book Co., Inc., New York, 1957, 488 pp. (see p. 414).

[*Received April 17, 1968; revised May 29, 1968*]